Robotics

L07: Perception (Part 2)

Dr. Mohammed Elmogy

Mansoura University
Faculty of Computers & Information Sciences
Department of Information Systems

elmogy@gmail.com

10. December 2011

Outline

- Active Ranging Sensors (P2)
  - Laser Rangefinder (Time-of-Flight, Electromagnetic)
  - 3D Laser Range Finder
  - 3D Range Sensor: Time Of Flight (TOF) camera
  - Triangulation Ranging
- Vision
  - Overview
  - Pinhole Model
  - Stereo Vision

Outline

- Active Ranging Sensors (P2)
  - Laser Rangefinder (Time-of-Flight, Electromagnetic)
  - 3D Laser Range Finder
  - 3D Range Sensor: Time Of Flight (TOF) camera
  - Triangulation Ranging
- Vision

Laser Rangefinder (Time-of-Flight, Electromagnetic)

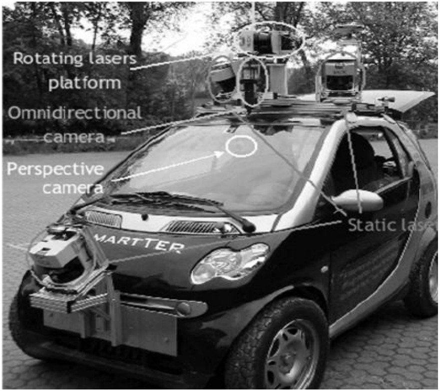
- It achieves significant improvements over the ultrasonic range sensor owing to the use of laser light instead of sound.
- It is often referred to as optical radar or lidar (**LIght Detection And Ranging**).
- These devices produce a range estimate based on the time needed for the light to reach the target and return.
- A mechanical mechanism with a mirror sweeps the light beam to cover the required scene in a plane (2D) or even in three dimensions, using a rotating, nodding mirror (3D).

Laser Rangefinder (cont.)

SICK

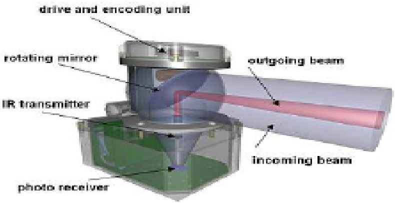
Alaska-IBEO

Hokuyo



Laser Rangefinder (cont.)

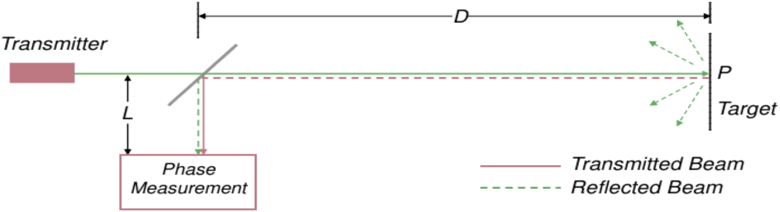
- This type of sensor consists of:
  - A transmitter which illuminates a target with a collimated beam (e.g., laser).
  - A receiver capable of detecting the component of light which is essentially coaxial with the transmitted beam.
  - Receiver detects the time needed for round-trip.



Laser Rangefinder (cont.)

- Operating Principles:
  - Use a pulsed laser and then measure the elapsed time directly, just as in the ultrasonic. (capable of resolving picoseconds) (today the standard).
  - Measure the beat frequency between a frequency-modulated continuous wave (FMCW) and its received reflection.
  - Measure the phase shift of the reflected light.

Phase-Shift Measurement



$$D' = L + 2D = L + \frac{\theta}{2\pi}\lambda \qquad \lambda = \frac{c}{f}$$

Where:

c: is the speed of light; *f* the modulating frequency; *D'* the distance covered by the emitted light is.

- for f = 5 MHz (as in the A.T&T. sensor), λ = 60 meters

Phase-Shift Measurement (cont.)

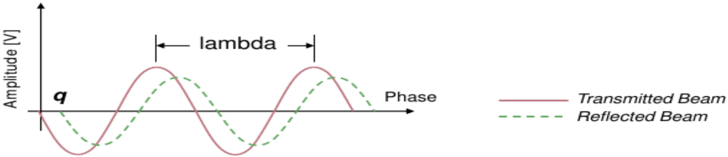
- Near-infrared light from a light-emitting diode [LED] or laser.
- Diffuse reflection will occur for surfaces having a roughness greater than the wavelength of the incident light.
- The wavelength of the infrared light emitted is 824 nm and so most surfaces, with the exception of only highly polished reflecting objects, will be diffuse reflectors.
- The sensor transmits 100% amplitude modulated light at a known frequency and measures the phase shift between the transmitted and reflected signals.

Phase-Shift Measurement (cont.)

- Distance D, between the beam splitter and the target

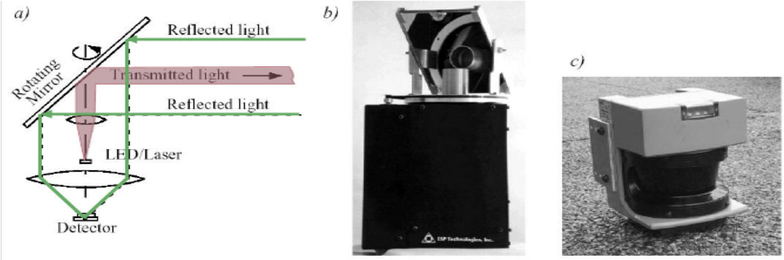
$$D = \frac{\lambda}{4\pi}\theta$$

- where
  - θ: phase difference between transmitted and reflected beam
- Theoretically ambiguous range estimates
  - since for example if $\lambda = 60$ meters, a target at a range of 5 meters = target at 35 meters



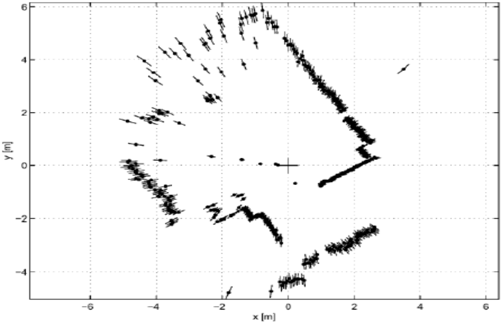
Phase-Shift Measurement (cont.)

- Uncertainty of the range (phase/time estimate) is inversely proportional t the square of the received signal amplitude.
    - Hence dark, distant objects will not produce such good range estimated as closer brighter objects …



Phase-Shift Measurement (cont.)

- Typical range image of a 2D laser range sensor with a rotating mirror. The length of the lines through the measurement points indicate the uncertainties.



The SICK LMS 200 Laser Scanner

- Angular resolution 0.25 deg
- Depth resolution ranges between 10 and 15 mm and the typical accuracy is 35 mm, over a range from 5 cm up to 20 m or more (up to 80 m), depending on the reflectivity of the object being ranged.
- This device performs seventy five 180-degrees scans per second

Note 1

Laser rangefinders cannot detect the presence of optically transparent materials such as glass, and this can be a significant obstacle in environments.

Note 2

As with ultrasonic ranging sensors, an important error mode involves coherent reflection of the energy. With light, this will only occur when striking a highly polished surface.

Note 1

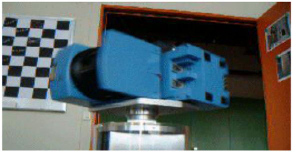
Laser rangefinders cannot detect the presence of optically transparent materials such as glass, and this can be a significant obstacle in environments.

Note 2

As with ultrasonic ranging sensors, an important error mode involves coherent reflection of the energy. With light, this will only occur when striking a highly polished surface.

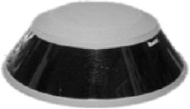
3D Laser Range Finder

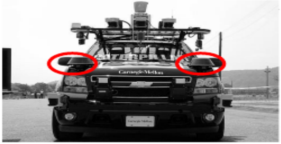
- **A 3D laser range finder is a laser scanner that acquires scan data in more than a single plane.**
- Custom-made 3D scanners are typically built by nodding or rotating a 2D scanner in a stepwise or continuous manner around an axis parallel to the scanning plane.
- By lowering the rotational speed of the turn-table, the angular resolution in the horizontal direction can be made as small as desired.
- A full spherical field of view can be covered (360° in azimuth and +/-90° in elevation).
- **However, acquisition takes up to some seconds!**
  For instance, if our laser takes 75 plane-scans/sec and we need an azimuthal angular resolution of 0.25 degrees, the period for a half rotation of the turn-table necessary to capture a spherical 3D scan with two Sicks is then 360 / 0.25 / 75 / 2 = 9.6 seconds. If one is satisfied with an azimuthal angular resolution of 1 degree, then the acquisition time drops down to 2.4 seconds, which is still too high for 3D mapping during motion!



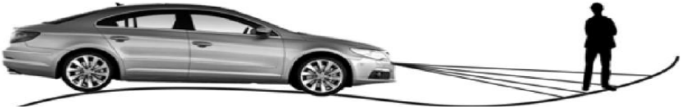
3D Laser Range Finder (cont.)

- The **Alasca XT laser scanner** splits the laser beam into **four vertical layers** with an aperture angle of 3.2°.
- This sensor is typically used for obstacle and pedestrian detection on cars. Because of its multi-layer scanning principle, it allows us any pitching of the vehicle



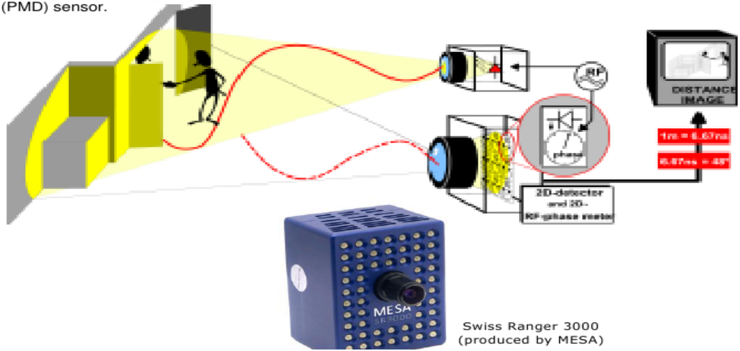


C Carnegie Mellon University



3D Range Sensor: Time Of Flight (TOF) camera

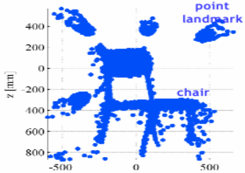
- A Time-of-Flight camera (TOF camera, figure) works similarly to a lidar with the advantage that **the whole 3D scene is captured at the same time and that there are no moving parts.** This device uses a modulated infrared lighting source to determine the distance for each pixel of a Photonic Mixer Device (PMD) sensor.



Swiss Ranger 3000
(produced by MESA)

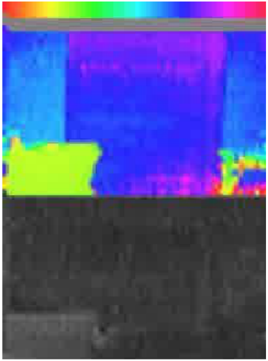
Incremental Object Part Detection

- Range Camera
  - 3D information with high data rate (100 Hz)
  - Compact and easy to manage
  - High, non-uniform measurement noise
  - High outlier rate at jump edges
  - However very low resolution (174x144 pixels)

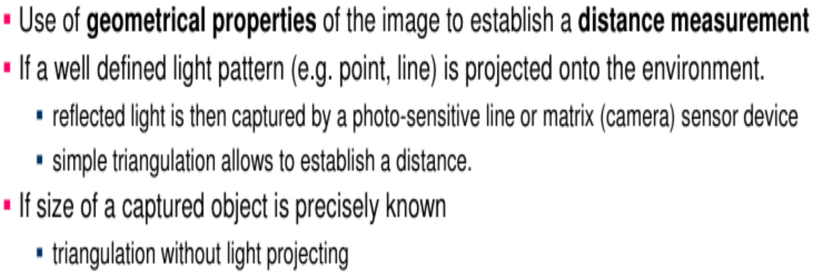


3D Point Cloud

Range Camera SR-3000

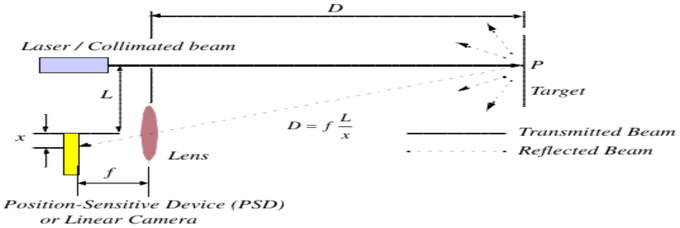


Triangulation Ranging



- Use of **geometrical properties** of the image to establish a **distance measurement**
- If a well defined light pattern (e.g. point, line) is projected onto the environment.
  - reflected light is then captured by a photo-sensitive line or matrix (camera) sensor device
  - simple triangulation allows to establish a distance.
- If size of a captured object is precisely known
  - triangulation without light projecting

Laser Triangulation (1D)



- Principle of 1D laser triangulation: $D = f\,\frac{L}{x}$

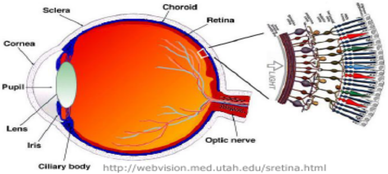
Outline

Active Ranging Sensors (P2)
Vision
- Overview
- Pinhole Model
- Stereo Vision

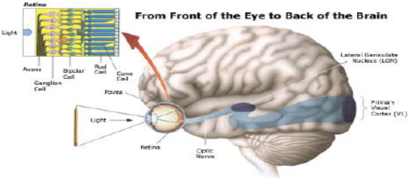
One picture, a thousand words

- Of all our senses, **vision** is the most powerful in aiding our perception of the 3D world around us.
- Retina is ~1000m^2. Contains millions of **photoreceptors**
  (120 mil. rods and 7 mil. Cones for colour sampling)
- Provides **enormous** amount of information: data-rate of ~3GBytes/s

➩ a large proportion of our brain power is dedicated to processing the signals from our eyes



http://webvision.med.utah.edu/sretina.html


Retina
Light
Axons
Ganglion Cell
Bipolar Cell
Rod Cell
Cone Cell
From Front of the Eye to Back of the Brain
Lateral Geniculate Nucleus (LGN)
Fovea
Light
Primary Visual Cortex (V1)
Retina
Optic Nerve

Human Visual Capabilities

- Our visual system is very sophisticated
- Humans can interpret images successfully under a wide range of conditions – even in the presence of very limited cues

Vision for Robotics

- Enormous descriptability of images
  ➪ a lot of data to process (human vision involves 60 billion neurons!)
- Not sensible to copy the biology, but learn from it
- Capture light ➪ Convert to digital image
  ➪ Process to get 'salient' information
- Vision is increasingly popular as a sensing modality:
  - compactness,
  - compatibility,
  - low cost, …
  - HW advances necessary to support the processing of images

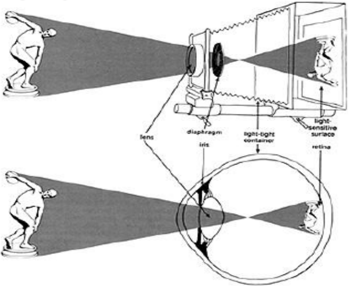
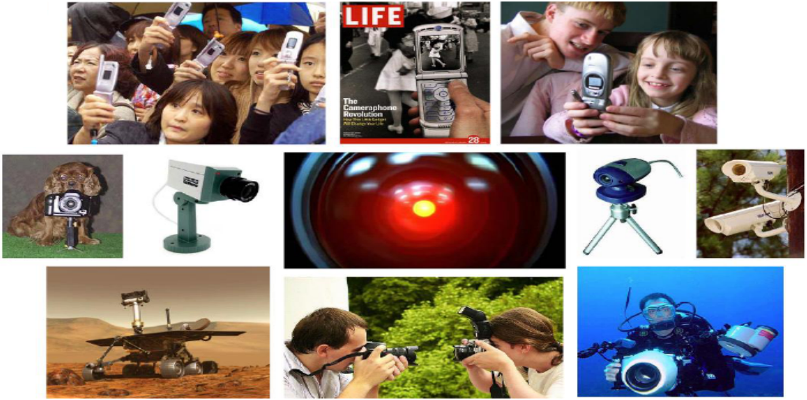
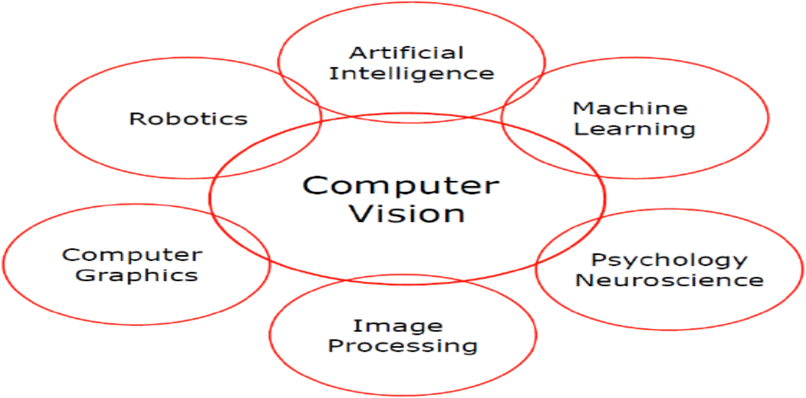


Image: Nicholas M. Short

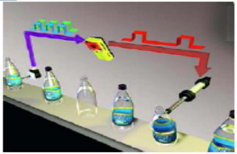
Computer Vision



Connection to other disciplines

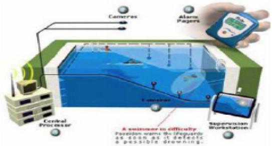


Applications of Computer Vision



Factory inspection

Reading license plates, checks, ZIP codes



Monitoring for safety (Poseidon)

Surveillance

Autonomous driving, robot navigation

Driver assistance
(collision warning, lane departure warning, rear object detection)

Applications of Computer Vision

Assistive technologies

Entertainment
(Sony EyeToy)

Movie special effects

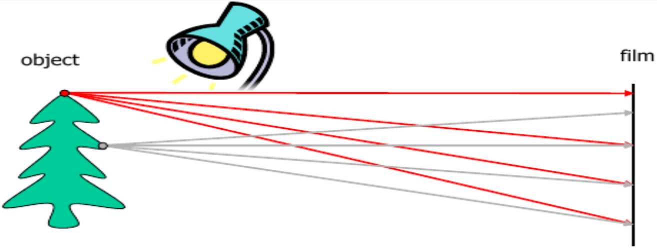
Canon
[Face priority AE] When a bright part of the face is too bright

Digital cameras (face detection for setting focus, exposure)

Visual search
http://www.kooaba.com/

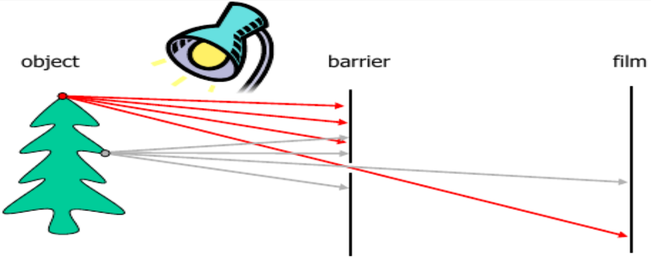
The camera

How do we see the world?



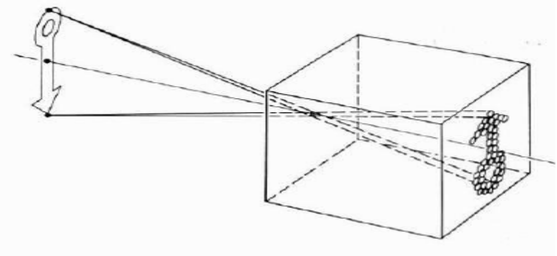
- Place a piece of film in front of an object
  ⇨ Do we get a reasonable image?

Pinhole camera



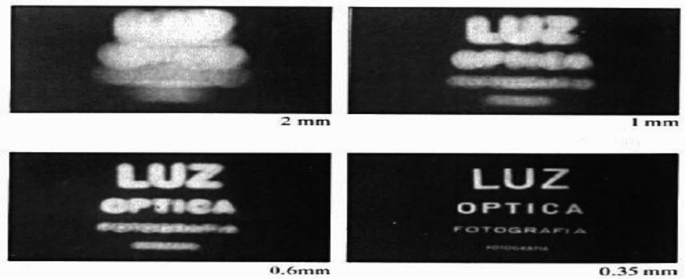
- Add a barrier to block off most of the rays
  - This reduces blurring
  - The opening known as the **aperture**

Pinhole camera model



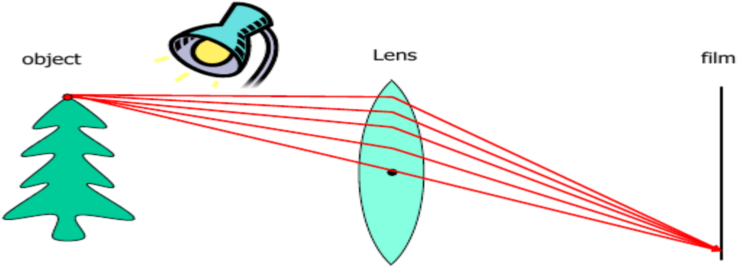
- Pinhole model:
  - Captures **pencil of rays** – all rays through a single point
  - The point is called **Center of Projection** or **Optical Center**
  - The image is formed on the **Image Plane**

Shrinking the aperture



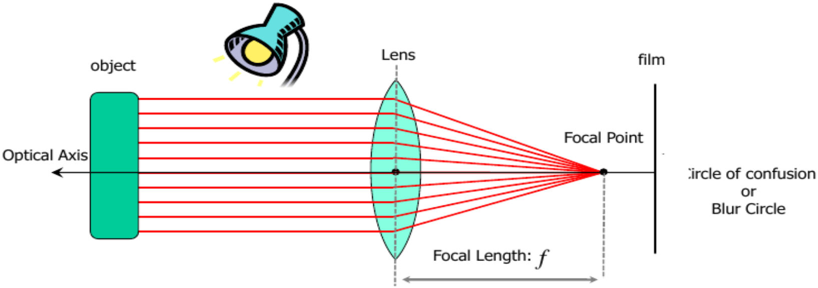
- Why not make the aperture as small as possible?
    - Less light gets through (must increase the exposure)
    - Diffraction effects…

Solution: adding a lens



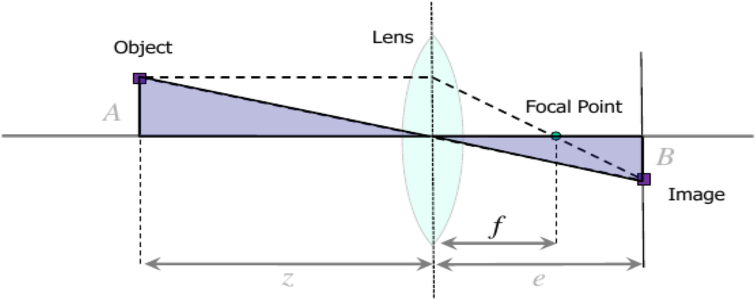
- A lens focuses light onto the film
  - Rays passing through the center are not deviated

Solution: adding a lens (cont.)



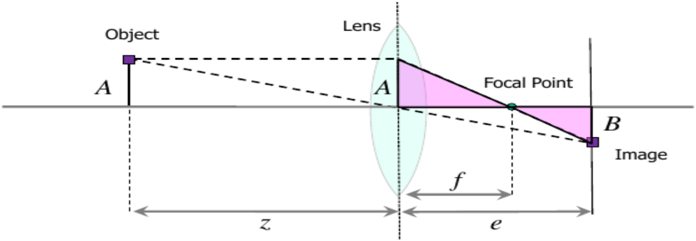
- A lens focuses light onto the film
  - Rays passing through the center are not deviated
  - All rays parallel to the **Optical Axis** converge at the **Focal Point**

Thin lens equation



- Similar Triangles: $\frac{B}{A} = \frac{e}{z}$

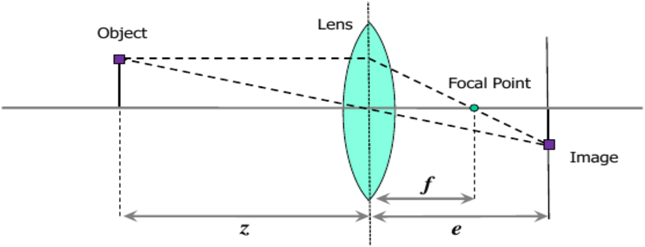
Thin lens equation (cont.)



- Similar Triangles:

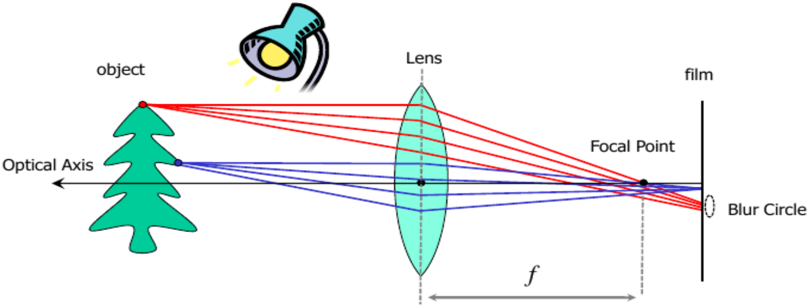
$$\left.\begin{array}{l}\dfrac{B}{A}=\dfrac{e-f}{f}=\dfrac{e}{f}-1\\[2ex]\dfrac{B}{A}=\dfrac{e}{z}\end{array}\right\}\ \dfrac{e}{f}-1=\dfrac{e}{z}\Rightarrow\boxed{\dfrac{1}{f}=\dfrac{1}{z}+\dfrac{1}{e}}$$

Thin lenses



- Thin lens equation: $\frac{1}{f} = \frac{1}{z} + \frac{1}{e}$
  - Any object point satisfying this equation is in focus
  - This formula can also be used to estimate roughly the distance to the object (“Depth from Focus”)

In Focus

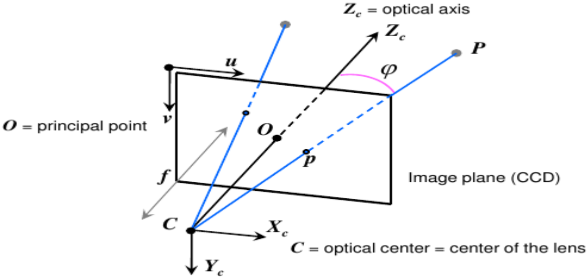


- There is a specific distance at which objects are “in focus”
- other points project to a “blur circle” in the image

Perspective

- The dependence of the apparent size of objects on their distance from the observer is known as perspective.
- Perspective gives us very strong depth cues $\Longrightarrow$ hence we can perceive a 3D scene by viewing its 2D representation (i.e. image).
- When viewing 3D scenes, it is possible to be mislead by perception.

Perspective Camera



- For convenience, the image plane is usually represented in front so that the image preserves the same orientation (i.e. not flipped)
- **Notice: a camera does not measure distances but angles! Therefore it is a "bearing sensor"**

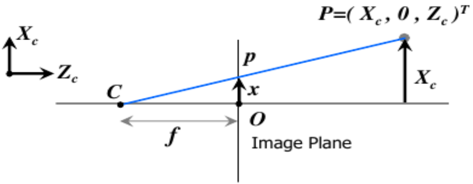
Perspective Projection

From the Camera frame to the local image plane

- The 3D scene point $P = (X_c \quad Y_c \quad Z_c)^T$ projects to point $p = (x \quad y)^T$ onto the (local) image plane

- Analysing similar triangles:

$$\frac{x}{f} = \frac{X_c}{Z_c} \Rightarrow x = \frac{fX_c}{Z_c}$$

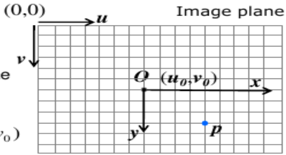


- So, for the 3D case, we can also obtain:

$$\frac{y}{f} = \frac{Y_c}{Z_c} \Rightarrow y = \frac{fY_c}{Z_c}$$

Perspective Projection (General)

From the Camera frame to pixel coordinates



- To convert the coordinates of $\boldsymbol{p}$, from the local image plane coordinates (x,y) to the pixel coordinates (u,v) we have to take account for:
  - Pixel coordinates of the camera optical center $O = (u_0, v_0)$
  - Scale factors for the pixel-size in both dimensions k_u, k_v

So:

$$u = u_0 + k_u x \Rightarrow u = u_0 + \frac{k_u f X_c}{Z_c}$$

$$v = v_0 + k_v y \Rightarrow v = v_0 + \frac{k_v f Y_c}{Z_c}$$

- Use **Homogeneous Coordinates** for linear mapping from 3D to 2D, by introducing an extra element:

$$p = \begin{pmatrix} u \\ v \end{pmatrix} \Rightarrow \tilde{p} = \begin{pmatrix} \lambda u \\ \lambda v \\ \lambda \end{pmatrix}$$

and similarly for the world coordinates. Note, usually $\lambda = 1$

Perspective Projection

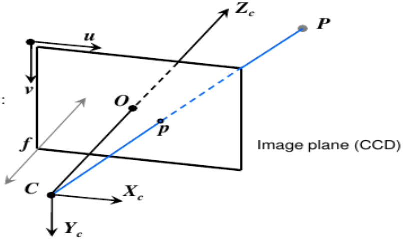
- So:

$$u = u_0 + \frac{k_u f X_c}{Z_c}$$

$$v = v_0 + \frac{k_v f Y_c}{Z_c}$$

Expressed in matrix form & Homogenerous coords:

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \begin{bmatrix} k_u f & 0 & u_0 \\ 0 & k_v f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$



Or alternatively

$$\begin{bmatrix} \lambda u \\ \lambda v \\ \lambda \end{bmatrix} = \underbrace{\begin{bmatrix} \boxed{\alpha_u} & 0 & u_0 \\ 0 & \boxed{\alpha_v} & v_0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Matrix of Intrinsic Parameters}} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = K \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$$

Focal length in u-direction

Focal length in v-direction

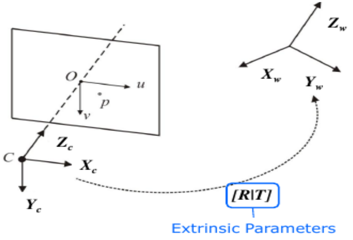
Projection onto the image plane

From the Camera frame to the World frame

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix}$$

in homogeneous coords

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \left[\begin{array}{ccc|c} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \left[\begin{array}{ccc|c} & R & & T \\ \hline 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

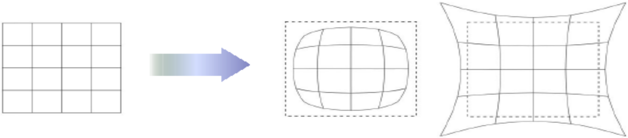


Projection Matrix

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \;\Longrightarrow\; \lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \boxed{K\left[R|T\right]} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

Radial distortion

- Straight lines in the world ➪ Straight lines in the image?



From ideal (u,v) to distorted pixel coordinates (u_d, v_d)

- Simple distortion model:

$$\begin{bmatrix}u_d\\v_d\end{bmatrix} = (1+\underbrace{k_1}_{\text{Radial Distortion parameter}} r^2)\begin{bmatrix}u-u_0\\v-v_0\end{bmatrix}+\begin{bmatrix}u_0\\v_0\end{bmatrix}$$

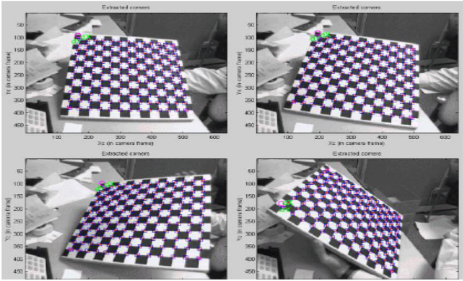
where

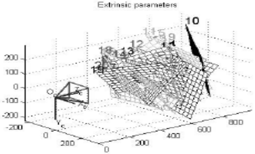
$$r^2 = (u-u_0)^2+(v-v_0)^2 .$$

Barrel distortion Pincushion distortion

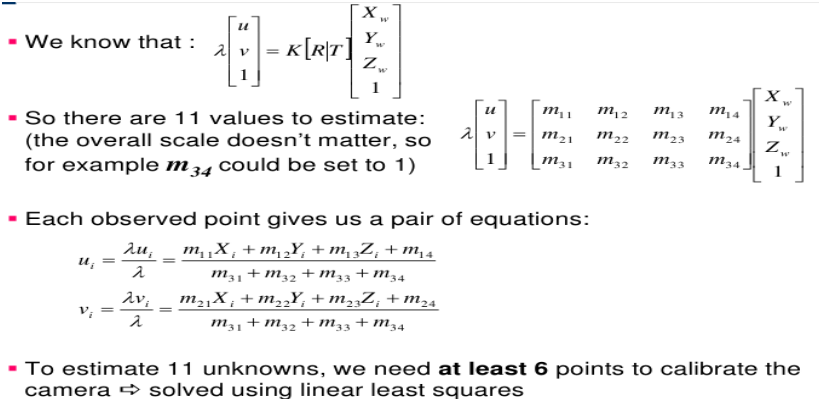
Camera Calibration

- Measure all unknown parameters to form the camera model (intrinsic & extrinsic) ⇨ use model to ‘understand’ the projection from world to image plane
- By knowing the pixel coordinates of the image points $\boldsymbol{p}$ and the 3D coordinates of the corresponding scene points $\boldsymbol{P}$, it is possible to compute the unknown parameters $\boldsymbol{K},\boldsymbol{R},\boldsymbol{T}$ and by solving the perspective projection equation





Camera Calibration (cont.)



- We know that : $\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$
- So there are 11 values to estimate: (the overall scale doesn't matter, so for example $\boldsymbol{m_{34}}$ could be set to 1) $\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$
- Each observed point gives us a pair of equations:

$$u_i = \frac{\lambda u_i}{\lambda} = \frac{m_{11}X_i + m_{12}Y_i + m_{13}Z_i + m_{14}}{m_{31} + m_{32} + m_{33} + m_{34}}$$

$$v_i = \frac{\lambda v_i}{\lambda} = \frac{m_{21}X_i + m_{22}Y_i + m_{23}Z_i + m_{24}}{m_{31} + m_{32} + m_{33} + m_{34}}$$

- To estimate 11 unknowns, we need **at least 6** points to calibrate the camera ⇨ solved using linear least squares

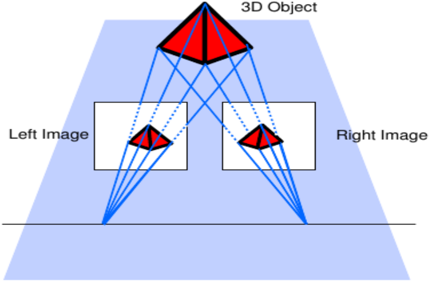
Camera Calibration (cont.)

- Having obtained the 3x4 projection matrix, it can be decomposed into the camera calibration matrix, the rotation and position of the camera: $K\left[R\middle|T\right]$
- Using standard techniques (QR decomposition) we can decompose the 3x3 submatrix into the product of an upper triangular matrix $\boldsymbol{K}$ and a rotation matrix $\boldsymbol{R}$ (orthogonal matrix)
- The translation $\boldsymbol{T}$ can subsequently be obtained by:

$$T = K^{-1}\begin{bmatrix} m_{14} \\ m_{24} \\ m_{34} \end{bmatrix}$$

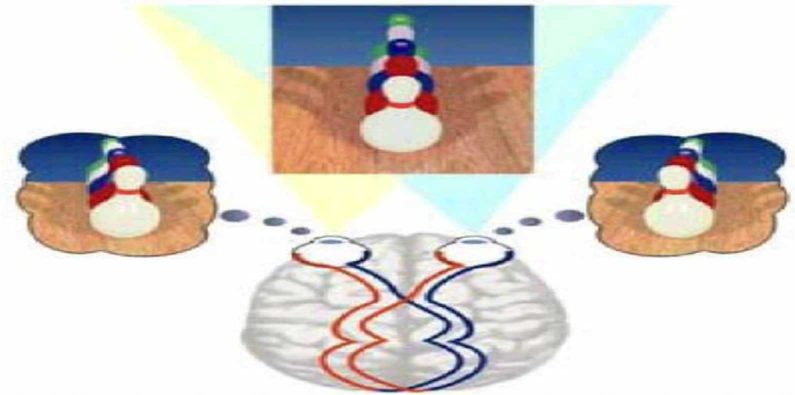
How do we measure distances with cameras?

- Impossible to capture 3D structure from a **single** image. We can only deduct the **ray** on which each image point lies.



- Observe scene from 2 different viewpoints ⇨solve for the intersection of the rays and recover the 3D structure

Disparity in the human retina



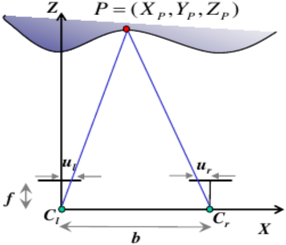
How do we measure distances with cameras?

- Structure from stereo (Stereo-vision):
  - use two cameras with known relative position and orientation

- Structure from motion:
  - use a single moving camera: both 3D structure and camera motion can be estimated up to a scale

Stereo Vision - The simplified case

- The simplified case is an ideal case. It assumes that both cameras are identical and are aligned on a horizontal axis



Baseline
distance between the optical centers of the two cameras

From Similar Triangles:

$$\frac{f}{Z_P}=\frac{u_l}{X_P}$$

$$\frac{f}{Z_P}=\frac{u_r}{b-X_P}$$

$$\Rightarrow \quad Z_P=\frac{bf}{u_l-u_r}$$

Disparity
difference in image location of the projection of a 3D point in two image planes

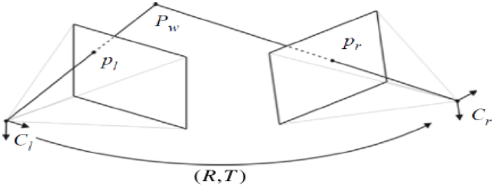
Stereo Vision facts

$$Z_P = \frac{bf}{u_l - u_r}$$

1. Depth is inversely proportional to disparity $(u_l - u_r)$
   - Foreground objects have bigger disparity than background objects
2. Disparity is proportional to stereo-baseline b
   - The smaller the baseline b the more uncertain our estimate of depth
   - However, as b is increased, some objects may appear in one camera, but not in the other (remember both cameras have parallel optical axes)
3. The projections of a single 3D point onto the left and the right stereo images are called **'correspondence pair'**

Stereo Vision – the general case

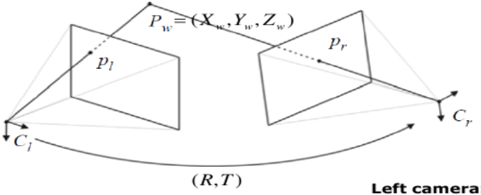
- Two identical cameras do not exist in nature!
- Aligning both cameras on a horizontal axis is very hard -- even in the most expensive stereo cameras!



- In order to be able to use a stereo camera, we need the
  - **relative pose** between the cameras (rotation, translation), and
  - the **focal length**, **image center**, **radial distortion** of each
  - ⇨ Use a calibration method, as mentioned before

Stereo Vision – the general case (cont.)

- To estimate the 3D position of P_w we just construct the system of equations of the left and right camera



Left camera:
set the world frame to coincide with the left camera frame

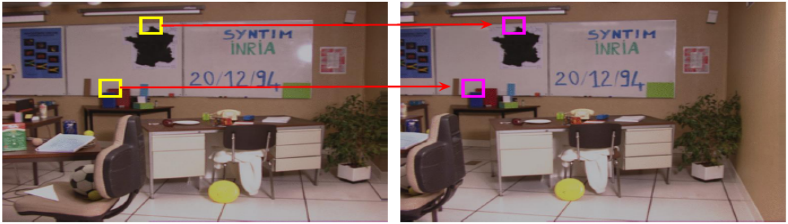
$$\tilde{p}_l = \lambda_l \begin{bmatrix} u_l \\ v_l \\ 1 \end{bmatrix} = K_l \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix}$$

Right camera:

$$\tilde{p}_r = \lambda_r \begin{bmatrix} u_r \\ v_r \\ 1 \end{bmatrix} = K_r R \begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} + T$$

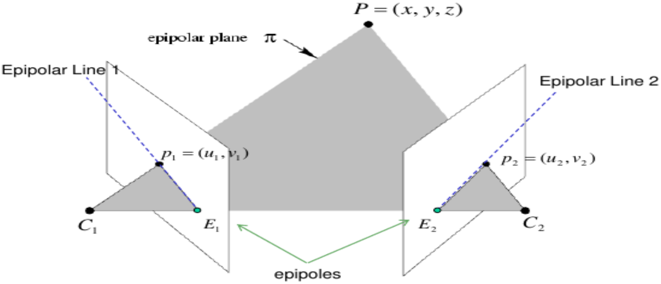
Correspondence Problem

- Which patch in the Left image, corresponds to the projection of the same 3D scene point on the Right image?
- Correspondence search could be done by testing the query-patches at all pixel positions in the other image. Typical **similarity measures** are the Cross-Correlation and Sum of Squared Differences
- **Exhaustive** image search can be computationally very expensive! Is there a way to make the correspondence search 1-dimensional?



Epipolar Geometry

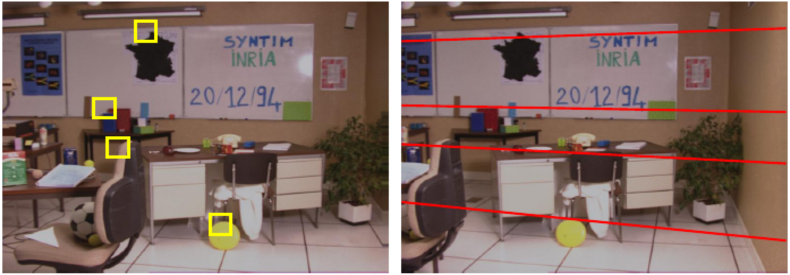
- Triangulating 2 rays from image correspondences
- Impose the **epipolar constraint** to aid matching: search for a correspondence along the epipolar line



- The epipolar plane is defined by a 3D point P and the optical centers.

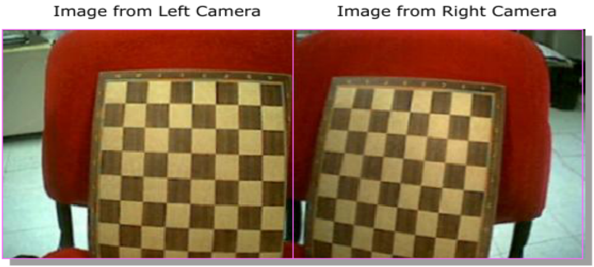
Correspondence Problem: Epipolar Constraint

- Thanks to the epipolar constraint, conjugate points can be searched along epipolar lines: this reduces the computational cost to 1 dimension!



Epipolar Rectification

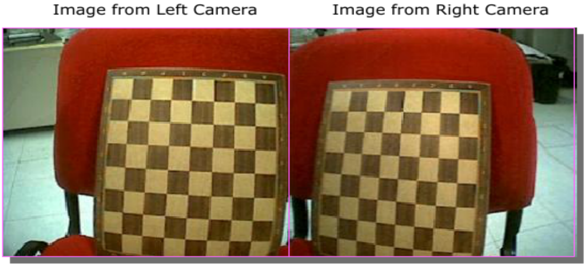
- Determines a transformation of each image plane so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal one)



Epipolar Rectification (cont.)

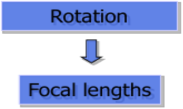
- Determines a transformation of each image plane so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal one)

Rotation



Epipolar Rectification (cont.)

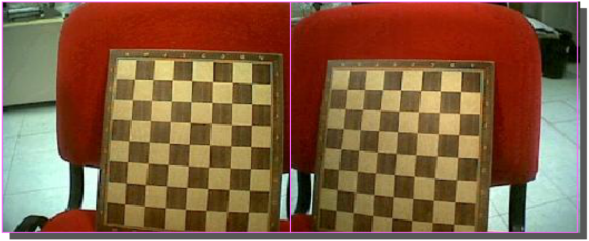
- Determines a transformation of each image plane so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal one)


Rotation

Focal lengths

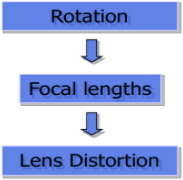
Image from Left Camera

Image from Right Camera



Epipolar Rectification (cont.)

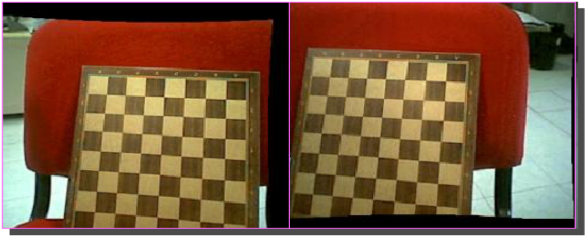
- Determines a transformation of each image plane so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal one)



Rotation
⇩
Focal lengths
⇩
Lens Distortion

Image from Left Camera

Image from Right Camera



Epipolar Rectification (cont.)

- Determines a transformation of each image plane so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal one)

Rotation
↓
Focal lengths
↓
Lens Distortion
↓
Translation

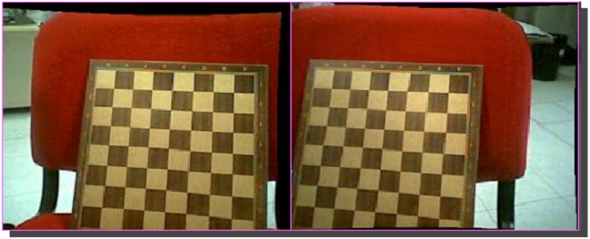
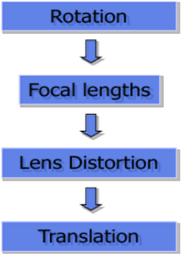


Image from Left Camera
Image from Right Camera

Epipolar Rectification (cont.)

- Determines a transformation of each image plane so that pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal one)

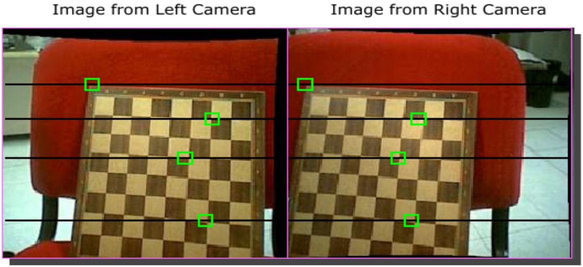


Rotation

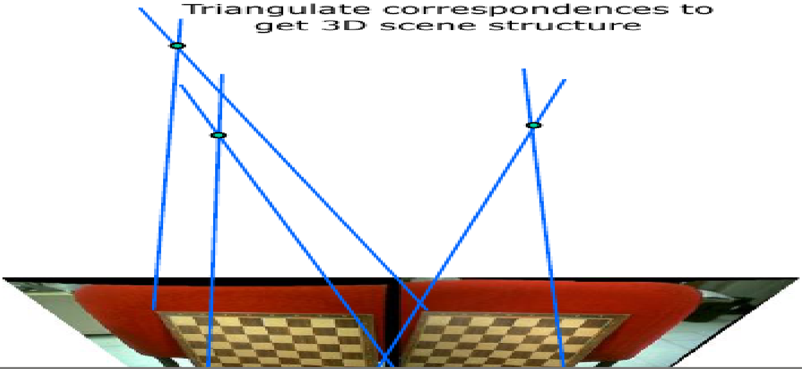
Focal lengths

Lens Distortion

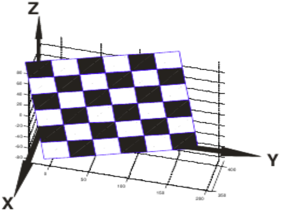
Translation

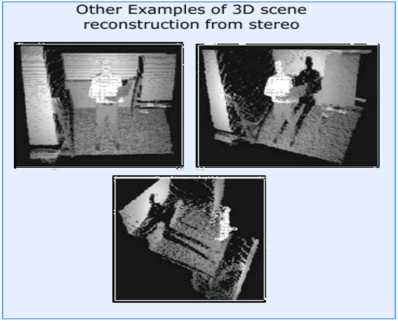


Stereo Vision Output 2 - 3D Reconstruction via triangulation



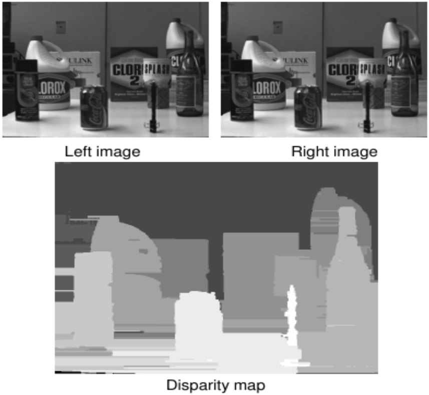
Stereo Vision Output 2 - 3D Reconstruction via triangulation (cont.)



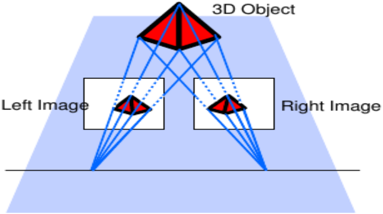

Other Examples of 3D scene reconstruction from stereo

Stereo Vision Output 1 – Disparity map

- Find the correspondent points of **all image pixels** of the original images
- Compute the disparity for each pair of correspondences
- **Disparity map**: holds the disparity value at every pixel
- Usually visualized as grey-scale images. If lighter color corresponds to larger disparities, then objects closer to the camera appear lighter, than those further away.



Stereo Vision - summary



1. Stereo camera calibration ⇨ compute camera relative pose
2. Epipolar rectification ⇨ align images & epipolar lines
3. Search for correspondences
4. Output: compute stereo triangulation or disparity map
5. Consider baseline and image resolution to compute accuracy